SUBRAMANIYAM VENKATA POONI

Chief Architect, Agentic Intelligence

Building Agentic LLM Systems with LangChain, Multi-Agent RAG, and Agent-Based Reasoning

AP2 | SLIM | X-A2A | LLMOps | Scalable AI Inference | Compilers & Runtimes: MLIR, LLVM, WASM

📧 svpooni70@yahoo.com | 📍 San Jose, CA 95135 | 🔗 linkedin.com/in/manip70 | 🇺🇸 US Citizen

EXECUTIVE SUMMARY

Visionary technology leader with 31+ years spanning AI/ML, distributed systems, cloud infrastructure, and enterprise architecture. Pioneer in Agentic AI, LLMOps, and HPC optimization. Track record: 6 US patents, 2 exits (Dell, IBM), $50M+ deals. Expert in building AI platforms for Fortune 500 and leading 500+ engineer teams.

KEY METRICS

31+	6	$50M+	2	12+	4
Years	Patents	Deals	Exits	Trade Secrets	Awards

PROFESSIONAL EXPERIENCE

CSSQUAREDB Technologies Inc.

Chief Architect & Founding Engineer, Agentic Intelligence

Sep 2025 – Present | San Jose, CA

- Multi-Stack GenAI on DGX/HGX B200, MI-350x, Nemotron, DeepSeek
- Agent Systems: LangChain, LangGraph, ReAct, Google ADK, Anthropic MCP
- Agentic RAG with A2A, AP2, SLIM protocols for enterprise LLM hosting

Broadcom (VMware)

Principal Architect – GenAI | HPC | SDDC

Dec 2021 – Aug 2025 | Palo Alto, CA

- PAIF-N AIaaS for CSPs: GPU-as-a-Service, AI PaaS, Model-as-a-Service (~40% margins)
- RAG Stack CI/CD, LLM Agent Programming (95%+ success, <2s latency, 10K+ calls/week)
- 3× MLPerf throughput, 40% latency reduction, 2× faster LoRA fine-tuning
- VMware Aria Automation 8.x driving $50M+ SOWs, 100% 5G Open RAN success

CSSQUAREDB Technologies

Founder CTO – Telco Cloud / 5G Open RAN

Dec 2020 – Dec 2021 | San Jose, CA

- DISH 5G Open RAN: 7.8K+ sites, Zero Touch Provisioning, 70% US coverage
- GitLab-Airflow Pipeline: 135+ TKG clusters, 1350 CNF in <6 hours

Personal Goal Pursuit (Career Break)

Key Projects & R&D

Sep 2019 – Dec 2020 | San Jose, CA

- Distributed ML Parameter Server (Python, Raft, Asyncio) – 40%+ throughput improvement
- MLIR-Based Compiler for WebAssembly AI Edge Inference (LLVM, MLIR)

Futurewei Technologies

Principal Lead – Software Competence Centre for Wireless, CTO Office

Feb 2016 – Sep 2019 | Santa Clara, CA

- SONiCS 1.0→4.0: Microservices, FaaS (1M+/day), MLaaS, Federated Learning
- Berkeley AI Research (BAIR) & Georgia Tech collaborations; 12+ trade secrets
- 🏆 Top Innovation Award | 🎖️ Future Star Medal | ⭐ WOW Team Award

A10 Networks, Inc.
Director of Research & Development, CTO Office
Dec 2011 – Jan 2016 | San Jose, CA

- 500+ engineers → DevOps with CI/CD & IaC, 3× faster releases
- SDN/NFV: Cisco ACI, VMware NSX, OpenStack ($10M+ savings)

Virtustream → Dell Acquisition
Principal SDN Architect
Aug 2010 – Dec 2011 | SF Bay Area

- xStream Platform: Storage/Networking for Data Centers; SDN/SDS orchestration lead

Hewlett Packard
Software Architect, Networking & Storage
Jul 1997 – Dec 2007 | Roseville, CA (10 yrs 6 mos)

- D2D Backup, OpenView Storage Area Manager (OVSAM); SNIA T10/T11 standards
- 6 US Patents in storage networking technology

IBM (Sequent Computers)
Software Engineer, Base Operating Systems
Mar 1994 – Jul 1997 | Beaverton, OR

- DYNIX/PTX OS kernel: thread scheduling, memory management, I/O for SMP/NUMA

TECHNICAL SKILLS

AI/ML & Agentic: LangChain, LangGraph, CrewAI, AutoGPT, OpenAI/Claude/Gemini API, HuggingFace, Pinecone, Weaviate, Chroma, FAISS, pgvector

NVIDIA/HPC: H100, A100, DGX B200, GH200, CUDA, TensorRT, Triton Server, NeMo, vLLM, RAPIDS, InfiniBand, BlueField DPU

DevOps/Cloud: Kubernetes, Docker, Terraform, Ansible, ArgoCD, GitLab CI, GitHub Actions, AWS, Azure, GCP, VMware VCF

Languages/Compilers: Python, Rust, Go, C/C++, Java, Scala | LLVM, MLIR, TVM, ONNX, WebAssembly, TensorRT

VMware Stack: VCF 5.2, Aria Automation 8.17, Aria Operations, NSX 4.x, vSAN 8, TKG, Cloud Director 10.6

US PATENTS

US 8,375,396 – Backup with Load Balancing
US 7,181,553 – Multiple Paths to SCSI Device
US 6,934,710 – Managing Fabric Device Access
US 7,610,295 – Persistent Path Identifiers
US 7,069,351 – SCSI Logical Unit IDs
US 10/260,419 – SCSI Devices on Linux

EDUCATION & CERTIFICATIONS

IIT Madras – Research Assistant, Computer Speech Lab (Premier Research Institution)

Master's – MIS, Mangalore University (First Class with Distinction) | **B.E. – CS&E**, SVCE (1987-1991)

Certifications: AI Agent Engineering (Maven 2025), VMware ACE/AAC 2025, VCF Livefire (NSX Federation, Aria Automation, Networking Design)

Courses: Functional Programming in Scala, Art of Functional Design (John De Goes) | Write a Compiler, Raft Consensus, Advanced Python (David Beazley)

AWARDS

🏆 Futurewei Top Innovation Award (GANs for Wireless, 2019) | 🎖️ Future Star Medal (Huawei, 2017) | ⭐ WOW Team Award (2018) | 🏅 Outstanding Contributions (2018)

KEY PROJECTS

- VMware Aria Automation | Multi-Cloud IaC – $50M+ SOWs, 100% 5G success, 50+ CI/CD flows, 90%+ SLA
- LLMOps Frameworks | RAG Observability – Mistral-7B, Phi-2, Gemma finetuning, 30% consistency, 4× CI/CD
- AI Performance Engineering – 3× MLPerf on NVIDIA (B200, H100) & AMD (MI350X), 40% latency reduction
- Distributed ML Parameter Server (Python, Raft) | MLIR Compiler for WASM Edge | Lox Interpreter (Rust)